# Chapter 7: Moving Beyond Linearity

So far we have mainly focused on linear models.

Linear models are simple to describe and implement.

Advantage: interpretation/inference.

disadvantage: can have limited predictive performance because linearity is always an approximation.

Previously, we have seen we can improve upon least squares using ridge regression, the lasso, principal components regression, and more.

improvement obtained by reducing complexity of OLS $\Rightarrow$ lowering variance.

Still a linear model! Can only be improved so much.

Through simple and more sophisticated extensions of the linear model, we can relax the linearity assumption while still maintiaining as much interpretability as possible. → extensions of linear model.

we've seen this one.

1. Polynomial regression: adding extra predictors that are original variables raised to a power.

   e.g. cubic regression uses $X, X^2, X^3$ as predictors, e.g. $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$.

   + non-linear fit

   − large powers can lead to strange shapes (especially near the boundary).

2. Step functions: cut the range of a variable into K distinct regions to produce a categorical variable. Fit a a piecewise constant function to X.

3. Regression Splines: more flexible than polynomials + step functions (extends both).

   idea: cut the range of X into K distinct regions + fit polynomial within each region

   Polynomials are constrained so that they are smoothly joined.

4. Generalized additive models (GAM): extends above to deal w/ multiple predictors.

We will start w/ predicting Y on X ($p=1$) and extend to multiple.

Note: We can talk regression or classification, e.g. logistic regression: $P(Y=1|X) = \dfrac{\exp(\beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_d X^d)}{1 + \exp(\beta_0 + \beta_1 X + \dots + \beta_d X^d)}$.

# 1 Step Functions

Using polynomial functions of the features as predictors imposes a *global* structure on the non-linear function of $X$.

We can instead use *step-functions* to avoid imposing a global structure.

i.e. break range of $X$ into bins and fit a different constant in each bin.

details: (1) create cut points $c_1, \dots, c_K$ in the range of $X$.

(2) Construct $K+1$ new variables

$$\left.\begin{aligned} C_0(X) &= \mathbb{I}(X < c_1) \\ C_1(X) &= \mathbb{I}(c_1 \le X < c_2) \\ &\vdots \\ C_K(X) &= \mathbb{I}(c_K \le X) \end{aligned}\right\} \text{indicator functions "dummy variables"}$$

Note for any $x$, $C_0(x) + C_1(x) + \dots + C_K(x) = 1$. since $x$ must be in exactly 1 interval.

leave out $C_0(X)$ because it is equivalent to fitting an intercept.

(3) Use OLS to fit linear model using $C_1(X), \dots, C_K(X)$

$$Y = \beta_0 + \beta_1 C_1(X) + \dots + \beta_K C_K(X) + \varepsilon.$$

For a given value of $X$, at most one of $C_1, \dots, C_K$ can be non-zero. ($C_K = C_K(X)$)

When $X < c_1 \Rightarrow$ all predictors $C_1, \dots, C_K = 0$

$\Rightarrow \beta_0$ interpreted as mean value for $Y$ when $X < c_1$.

$\beta_j$ represents the average increase in mean response for $X \in [c_j, c_{j+1})$ relative to $X < c_1$.

We can also fit the logistic regression model for classification:

$$P(Y=1|X) = \frac{\exp(\beta_0 + \beta_1 C_1(X) + \dots + \beta_K C_K(X))}{1 + \exp(\beta_0 + \beta_1 C_1(X) + \dots + \beta_K C_K(X))}$$

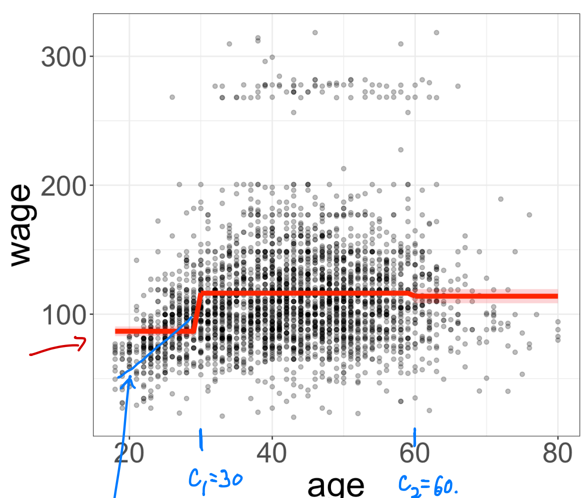
new interpretation of $\beta$'s related to log-odds.

Example: Wage data. n=3000 male workers in Mid-atlantic region.

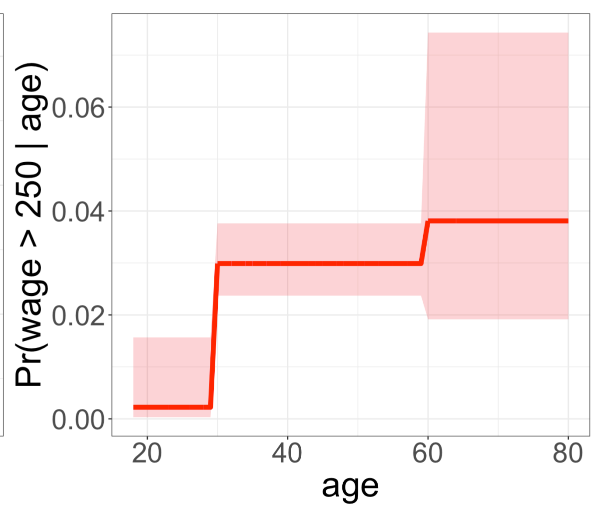
x y

| year | age | maritl | race | education | region | jobclass | health | health_ins | logwage | wage |
|---|---|---|---|---|---|---|---|---|---|---|
| 2006 | 18 | 1. Never Married | 1. White | 1. < HS Grad | 2. Middle Atlantic | 1. Industrial | 1. <=Good | 2. No | 4.318063 | 75.04315 |
| 2004 | 24 | 1. Never Married | 1. White | 4. College Grad | 2. Middle Atlantic | 2. Information | 2. >=Very Good | 2. No | 4.255273 | 70.47602 |
| 2003 | 45 | 2. Married | 1. White | 3. Some College | 2. Middle Atlantic | 1. Industrial | 1. <=Good | 1. Yes | 4.875061 | 130.98218 |
| 2003 | 43 | 2. Married | 3. Asian | 4. College Grad | 2. Middle Atlantic | 2. Information | 2. >=Very Good | 1. Yes | 5.041393 | 154.68529 |

fitted value of wage using step function of age.




Missed increasing trend here.




logistic regression modeling prob. of being a high earner (wage > 250k).

Using step function w/ knots at x=30,60.

Unless there are natural breakpoints in the predictor, piecewise constant functions can miss trends.

## 2 Basis Functions

Polynomial and piecewise-constant regression models are in fact special cases of a *basis function approach.*

**Idea:**

have a family of functions or transformations that can be applied to a variable $X$

$b_1(X), b_2(X), \dots, b_K(X).$

Instead of fitting the linear model in $X$, we fit the model

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \dots + \beta_K b_K(x_i) + \varepsilon_i$$

Note that the basis functions are fixed and known. (we choose them).

e.g. Polynomial regression: $b_j(x_i) = x_i^j$, $j = 1, \dots, d$

e.g. step functions: $b_j(x_i) = \mathbb{I}(c_j \le x_i < c_{j+1})$.

We can think of this model as a standard linear model with predictors defined by the basis functions and use least squares to estimate the unknown regression coefficients.

$\Rightarrow$ We can use all our inference tools for linear models, e.g. $se(\hat{\beta}_j)$ and F-statistics for model significance.

Many alternatives exist for basis functions:

e.g. Wavelets, Fourier series, regression splines (next).

# 3 Regression Splines

*Regression splines* are a very common choice for basis function because they are quite flexible, but still interpretable. Regression splines extend upon polynomial regression and piecewise constant approaches seen previously.

start with

## 3.1 Piecewise Polynomials

Instead of fitting a high degree polynomial over the entire range of $X$, piecewise polynomial regression involves fitting separate low-degree polynomials over different regions of $X$.

single cutpoint.

i.e. fit two different polynomials to data: 1 on subset for $x < c$ and a second on subset for $x \geq c$.

For example, a pieacewise cubic with no knots is just a standard cubic polynomial.

A pieacewise cubic with a single knot at point $c$ takes the form

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \varepsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \varepsilon_i & \text{if } x_i \geq c \end{cases}$$

each polynomial can be fit using least squares.

Using more knots leads to a more flexible piecewise polynomial.

If we place $L$ knots $\Rightarrow$ fit $L+1$ polynomials (doesn't have to be cubic).

In general, we place $L$ knots throughout the range of $X$ and fit $L + 1$ polynomial regression models.

This leads to $(d+1)(L+1)$ degrees of freedom in the model

(# parameters to fit $\approx$ complexity/flexibility).

## 3.2 Constraints and Splines

To avoid having too much flexibility, we can *constrain* the piecewise polynomial so that the fitted curve must be continuous.

i.e. there cannot be a jump at the knots.

To go further, we could add two more constraints

(1) first derivatives of the piecewise polynomials are continuous at the knots

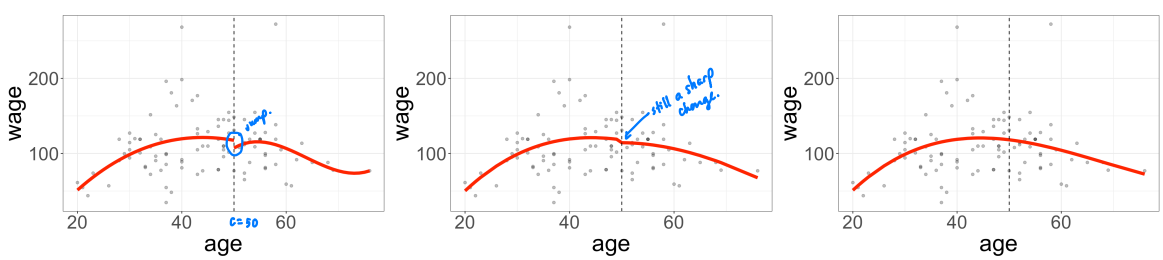
(2) $2^{nd}$ derivatives of the piecewise polynomials are continuous at the knots.

In other words, we are requiring the piecewise polynomials to be *smooth.*

Each constraint that we impose on the piecewise cubic polynomials effectively frees up one degree of freedom, by reducing the complexity of the resulting fit.

The fit with continuity and smoothness contraints is called a *spline*.

A degree-$d$ spline is a piecewise degree-$d$ polynomial w/ continuity in derivatives up to degree $d-1$ at each knot.




piecewise cubic polynomial.

piecewise cubic with continuity

cubic spline
cts + cts $1^{st}$ and $2^{nd}$ derivatives (smooth).

## 3.3 Spline Basis Representation

Fitting the spline regression model is more complex than the piecewise polynomial regression. We need to fit a degree $d$ piecewise polynomial and also constrain it and its up to $d-1$ derivatives to be continuous at the knots.

We can use the basis model to represent a regression spline.

e.g. cubic spline w/ L knots:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \dots + \beta_{L+3} b_{L+3}(x_i) + \varepsilon_i$$

for appropriate functions $b_1, \dots, b_{L+3}$.

$x, x^2, x^3$

The most direct way to represent a cubic spline is to start with the basis for a cubic polynomial and add one *truncated power basis* function per knot.

$$h(x, c) = (x-c)_+^3 = \begin{cases} (x-c)^3 & \text{if } x > c \\ 0 & \text{o.w.} \end{cases}$$ where $c$ is the knot.

$$\Rightarrow \quad y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \sum_{j=1}^{L} \beta_{3+j} h(x_i, c_j) + \varepsilon_i$$

see homework → This will lead to discontinuity in only the 3rd derivative at each $c_j$ with continuous first and second derivatives and continuity at each $c_j$.

df : $L+4$ (cubic spline w/ L knots).

Unfortunately, splines can have high variance at the outer range of the predictors. One solution is to add *boundary constraints*.

i.e. X is small or large.

require function to be linear at the boundary (where X is smaller than the smallest knot or bigger than the biggest knot)

"natural spline"

additional constraint produces more stable estimates at the boundaries.

## 3.4 Choosing the Knots

When we fit a spline, where should we place the knots?

Regression spline is most flexible in regions that contain a lot of knots (coefficients can change more rapidly).
⟹ place knots where we think relationship changes rapidly (less stable).

More common in practice: place them uniformly
to do this, choose desired degrees of freedom (flexibility) + use software to automatically place knots at uniform quantiles of the data.

How many knots should we use?

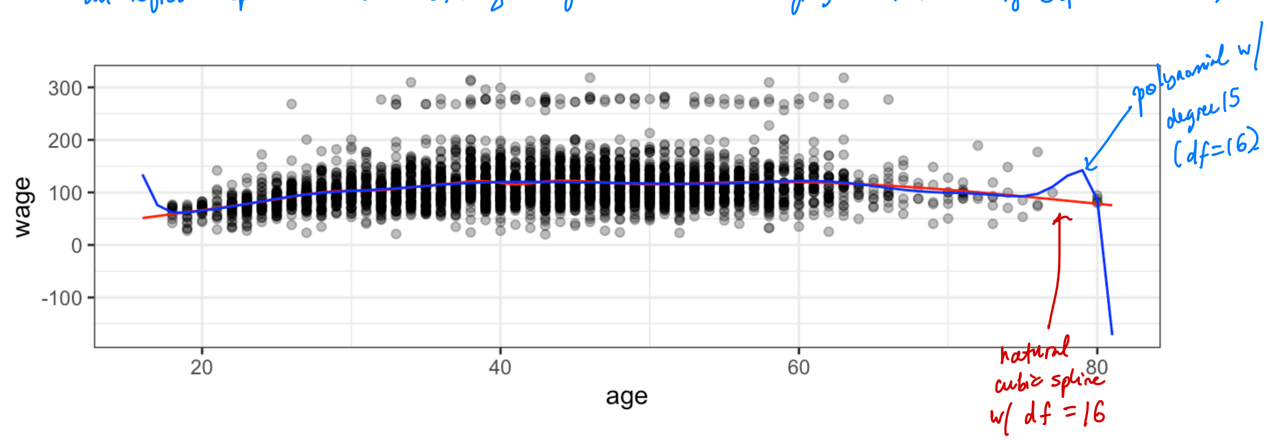
⟺ how many degrees of freedom should we use?

Use CV! Choose L that gives smallest CV error!

alternative: penalized splines (splines + lasso).

## 3.5 Comparison to Polynomial Regression

Regression splines often gives superior results to polynomial regression.
Polynomial regression must use high degree to achieve flexible fit (e.g. $X^{15}$),
but regression splines introduce flexibility through knots (but fixed degree) ⟹ more stability (esp. at boundaries).




extra flexibility of polynomial produces undesireable result at the boundary but spline w/ same flexibility still reasonable.

# 4 Generalized Additive Models

So far we have talked about flexible ways to predict $Y$ based on a single predictor $X$.

These approaches can be seen as extensions of simple linear regression
$Y = \beta_0 + \beta_1 X + \varepsilon$.

*Generalized Additive Models (GAMs)* provide a general framework for extending a standard linear regression model by allowing non-linear functions of each of the variables while maintaining *additivity*.

flexibly predicting $Y$ on the basis of several predictors $X_1, \dots, X_p$

## 4.1 GAMs for Regression

– still additive models
↳ can also be used for classification using logistic regression.

A natural way to extend the multiple linear regression model to allow for non-linear relationships between feature and response:

linear regression: $y_i = \beta_0 + \beta_1 x_{i1} + \dots + \beta_p x_{ip} + \varepsilon_i$

idea: replace each linear component $\beta_j x_{ij}$ with a smooth non-linear function:

$\Rightarrow$ GAM: $y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \varepsilon_i$

$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \dots + f_p(x_{ip}) + \varepsilon_i$

"additive" because we calculate a separate $f_j$ for each $X_j$ and add them together.

possibilities for $f_j$:

- identity (leads to linear regression).
- polynomial
- regression splines
- smoothing splines, local linear regression. → see textbook ch. 7.5–7.6

The beauty of GAMs is that we can use our fitting ideas in this chapter as building blocks for fitting an additive model.

Example: Consider the Wage data.

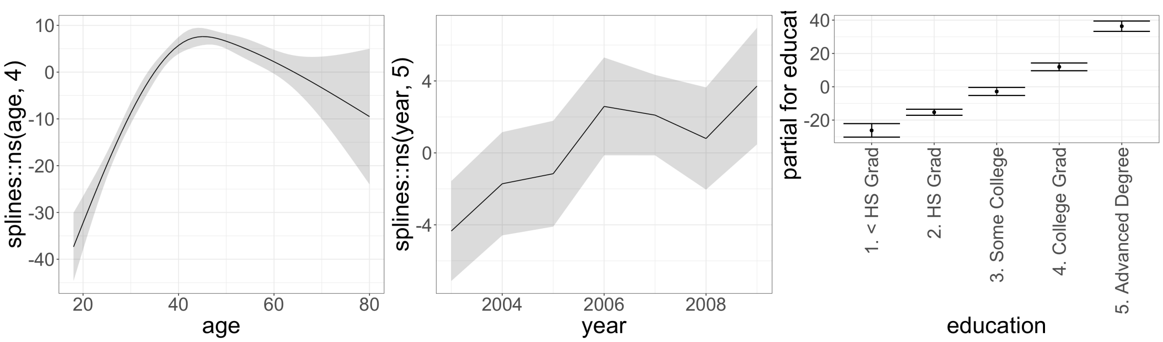
$$\text{Wage} = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education}) + \varepsilon$$

where $f_1$ is natural spline w/ 4 df

$f_2$ is natural spline w/ 5 df

$f_3$ is identity of dummy variables created from education.

easy to fit least squares by choosing appropriate basis functions.




Relationship btw/ each variable and the response (holding others fixed):

- age: hold year and education fixed, wage is low for young people and old people, highest for intermediate age.

- year: holding age and education fixed, wage tends to increase w/ year (inflation?)

- education: holding year & age fixed, wage increases w/ education.

We could easily replace $f_j$ w/ different smooth functions and get a different

just change the basis functions and use OLS.

## Pros and Cons of GAMs

Advantages:

- nonlinear fit $f_j$ to each $X_j$
- nonlinear fit can potentially lead to more accurate predictions (if truly nonlinear relationship).
- additive model $\Rightarrow$ can still interpret effect of each predictor (holding others fixed)

  $\Rightarrow$ useful for inference/interpretation.

Limitations:

- model is restricted to be additive

  i.e. important interactions can be missed

  Solution: we could manually add interaction terms by including additional predictors:

  - $X_j \cdot X_k$
  - low dimensional interaction function $f_{jk}(X_j, X_k)$

    $\uparrow$ two-dimensional spline (not covered).

GAMs provide a useful compromise btw linear and fully nonparametric models.

$\downarrow$ like random forests and boosted trees (next).

## 4.2 GAMs for Classification

assume $Y$ binary $(0,1)$

GAMs can also be used in situations where $Y$ is categorical. Recall the logistic regression model:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p$$

logit = log-odds of $P(Y=1|x)$ vs. $P(Y=0|x)$. = linear function of predictors

A natural way to extend this model is for non-linear relationships to be used.

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + f_1(X_1) + \dots + f_p(X_p)$$

logistic regression GAM.

Example: Consider the Wage data.

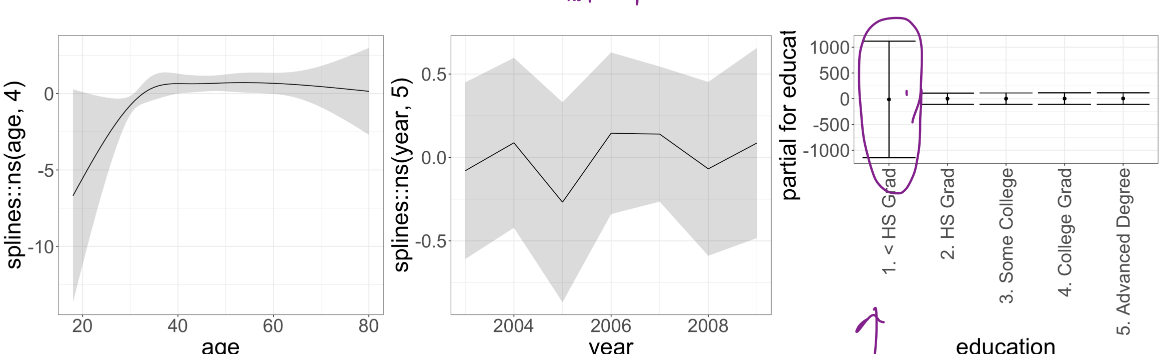
Let $Y$ = Wage > \$250k

we could fit a GAM

dummy variable, identity model.

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + f_1(\text{year}) + f_2(\text{age}) + f_3(\text{education})$$

df=4 natural splines df=5




increase in prob. up to 45, then levels off.

slight increase, not much.

Nobody in data set w/ < HS making more than \$250k